Appendix B. Study Methodology

At the initiation of this study, extensive consultation among the authors was conducted to discuss the advantages and disadvantages of various methods for characterizing fields of science and institutional boundaries, perceived challenges to assembly and analysis of the database, methods for addressing those challenges, preferred methods for interacting and study reporting, and proposed analysis approaches.

The following sections address key methodological issues of this study. Sections B.1-3 present information on issues concerned with defining and specifying the U.S. top 200 R&D performing institutions. Section B.4 discusses issues of consistency and availability of several independent variables. Section B.4 presents the sources of the data used in this study, and section B.5 discusses the methodological issues and preparation of the publication data for use in the study. Section B.6 explains how the Thomson ISI data were prepared in accordance with the agreed upon institutional groupings and field classifications. Section B.7 discusses the extraction and processing of NSF, NCES, and NRC data from the WebCASPAR database. Section B.8 provides information on database construction, quality control, and documentation. Section B.9 describes how the database was prepared, and how analyses and regressions were conducted.

B.1 Define Criteria for Membership in Top 200 Institutions

One of the critical issues addressed at the outset was how to define the top 200 U.S. academic institutions to be included in analyses. There were two possible criteria—publication outputs or R&D funding. One problem with using publication outputs is determining whether to use the fractional or whole count method alluded to in *Science and Engineering Indicators 2002*.[1] However, since the analyses used publication output as a dependent variable, it was decided to use an independent variable, R&D funding. Two sources of R&D funding were available. The Survey of Research and Development Expenditures at Universities and Colleges (academic R&D expenditures survey) measures actual R&D expenditures by year. The Federal (S&E) obligations data provide actual funding to the universities for a given year, but these data could be misleading because each obligation of funds could be for a variable period of one or many years even though it is obligated in one year. In addition, academic R&D expenditure data includes non-federal funding, which is a significant source of funding for universities. For this reason, it was decided to use the academic R&D expenditure survey data.

B.2 Define Institutional Boundaries of the Top 200 Institutions

In order to determine the 200 institutions with the highest average total academic R&D expenditures for 1988 through 2001, it was necessary to resolve institutional boundary issues as well as inconsistencies in how institutions reported total academic R&D expenditures data over time. Addressing institutional boundaries issues involves deciding whether to include branch campuses as part of an institution. Institutional boundaries within the WebCASPAR system were modified to assure compatibility both with the academic R&D expenditures data and ipIQ, Inc. (formerly CHI Research, Inc.) coding of institutions, and are shown in appendix A.

[1] National Science Board (Arlington, VA: National Science Foundation, 2002), NSB-02-1, p.5-40.

For systems or institutions with separate medical schools, the medical school was combined with the non-medical part of the system or institution. Other institutional boundaries issues required reviews of institutions' data.

Three institutions (John Hopkins University Applied Physics Laboratory, Woods Hole Oceanographic Institution, and the University of California President's Office) that were within the top 200 highest average total academic R&D expenditures for 1988 to 2001 were excluded from the database, because they had limited or no data for other data sources to be included in the database. Exclusion of these institutions resulted in three other institutions being included in the top 200.

B.3 Examine Feasibility of Defining a Residual Population of Institutions

Due to varying survey populations for the data sources to be included in the database, the residual population of institutions (those not in the top 200 U.S. academic institutions to be included in the database) differs for each data source. Defining a consistent residual population of institutions involves tracking a large number of individual institutions in each data source, and determining a set of institutions that consistently occurs in all data sources. Due to the level of effort and expense involved in defining a consistent residual population of institutions, a residual population of institutions was not included in the database.

B.4 Address Issues Related to Specific Independent Variables

Often resource and expenditure data were available at the departmental level while the outcome variables (publications and citations) were field-specific. Consequently, it was necessary to specify how to aggregate departmental data to the field level. We developed crosswalks between 3-digit department codes in the SED/DRF and a standardized set of fields used by ipIQ, Inc. (formerly CHI Research, Inc.) to classify publications and citations.

Carnegie classifications and definitions were significantly revised between 1994 and 2000. We decided to use the 1994 Carnegie classifications for all institutions, since that classification time was closer to the center of the time period under consideration (1988 to 2001).

The NCES IPEDS Finance Survey underwent significant changes after 1996. Although information from this survey was included in the database, and in some preliminary analyses, the inconsistency in longitudinal definitions prompted NSF to later decide to exclude these data from final analyses.

We had originally intended to analyze S&E Ph.D.s employed at an institution using data from the Survey of Doctorate Recipients (SDR). ORC Macro obtained the raw data files and examined detailed information on respondents concerning their place of employment, and included this information in the database. However, the survey was designed to ensure nationally representative estimates, rather than to provide precise estimates at the institutional level. The sample size at the institutional level is too small to make precise estimates. Partially for this reason, and partially because it was uncertain in what capacity these Ph.D.s were employed, NSF later decided not to include this variable in final analyses.

Institutions in the patents data were only identified by institution name. Therefore, it was necessary to develop a crosswalk of the institution names in the patents data to institution FICE codes. The crosswalk of patents data institution names to FICE codes was developed through

automated and manual comparisons of institution names. Although patents data are available by year, due to the sparseness of the data by year, all data for 1988 to 2001 were aggregated.

B.5 Extract Data for the Database's Population of Institutions

To construct the Publication Trends database, data were extracted from the following sources:

- Thomson ISI, Science Citation Index (SCI) and Social Sciences Citation Index (SSCI); data provided by ipIQ, Inc. (formerly CHI Research, Inc.).
- NCES IPEDS Finance Survey; WebCASPAR database system input SAS data sets.
- NCES IPEDS Opening Fall Enrollment Survey; WebCASPAR database system input SAS data sets.
- NCES IPEDS Salaries, Tenure, and Fringe Benefits of Full-time Instructional Faculty Survey; WebCASPAR database system input SAS data sets.
- NSF Survey of Doctorate Recipients; WebCASPAR database system, ABC-SQL Query Tool input data sets.
- NSF Survey of Earned Doctorates (SED)/Doctorate Records File (DRF); WebCASPAR database system input SAS data sets.
- NSF Survey of Research and Development Expenditures at Colleges and Universities; WebCASPAR database system input SAS data sets.
- NSF-NIH Survey of Graduate Students and Postdoctorates in Science and Engineering; WebCASPAR database system input SAS data sets.
- Research Doctorate Programs in the United States: Continuity and Change; National Academy of Sciences, National Research Council; WebCASPAR database system input SAS data sets.
- Patents data provided by the United State Patent and Trademark Office, Office of Electronic Information Products.

In addition, the following institution attributes were extracted from WebCASPAR source data sets for inclusion in the database:

- 1994 Carnegie Classification
- Type of Institutional Control (public/private)

B.6 Process SCI and SSCI Data

The Thomson ISI SCI and SSCI data were pre-aggregated by ipIQ, Inc. (formerly CHI Research, Inc.)[2] in accordance with the institutional boundaries used in the defining the 200 institutions included in the database. NSF provided ipIQ, Inc. with a detailed listing of institutional groupings by FICE code and institution name. This listing was prepared by ORC

[2] Thomson ISI, Science Citation and Social Sciences Citation Indexes, http:www.isinet.com/productis/citation/; ipIQ, Inc.; National Science Foundation (NSF), Division of Science Resources Statistics, special tabulations; and NSF WebCASPAR database, special tabulations.

Macro using the WebCASPAR crosswalk of child to parent institutions as well as the *Codebook for Compatible Statistical Reporting of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions* that includes some child institutions that are not in WebCASPAR, but may occur in the SCI/SSCI data. Also included for some institutional groupings were notes added by NSF to indicate institutions that have a medical school for which SCI and SSCI data should be included in the institutional grouping as well as to indicate other factors to consider in aggregating the data such as institutions to exclude, alternate institution names, and other institutional components (observatories, research institutes, etc.) to include.

For the Thomson ISI SCI and SSCI data, it was necessary to develop crosswalks of the institution and field identifiers used in those data to institution FICE codes and WebCASPAR field codes. The crosswalk of SCI/SSCI institution names to FICE codes was developed through automated and manual comparisons of institution names. Institution address information was also used to confirm mappings. The SCI/SSCI data contained field codes and names for 135 unique fields as classified by ipIQ, Inc. (formerly CHI Research, Inc.). These fields were mapped to WebCASPAR fields by manually comparing the name of each SCI/SSCI field to the survey-specific field names in the following WebCASPAR data sources for which crosswalks to WebCASPAR fields already existed:

- NSF Survey of Earned Doctorates (SED)/Doctorate Records File (DRF)
- NSF-NIH Survey of Graduate Students and Postdoctorates in Science and Engineering
- Research Doctorate Programs in the United States: Continuity and Change; National Academy of Sciences, National Research Council

Where possible, each SCI/SSCI field was mapped to a survey-specific field name in each of the above data sources. The existing crosswalks of survey-specific field to WebCASPAR field for each of these data sources were used to determine the appropriate WebCASPAR field for each SCI/SSCI field. Using this methodology, all but twelve SCI/SSCI fields were mapped to WebCASPAR fields. The remaining twelve SCI/SSCI fields were mapped to WebCASPAR fields through additional research of field definitions, including the use of a file provided by ipIQ, Inc. (formerly CHI Research, Inc.) that maps fields and subfields to journal titles. (For a crosswalk between WebCASPAR and ipIQ fields, see appendix E).

B.7 Process Data Extracted from WebCASPAR

NSF and NCES IPEDS data as well as doctorate program rating data from the National Research Council's (NRC) publication, *Research Doctorate Programs in the United States: Continuity and Change*, were extracted from WebCASPAR database system input SAS data sets prepared by ORC Macro. The NSF SDR data were extracted from the input SAS data set for WebCASPAR's ABC-SQL Query Tool component prepared by ORC Macro. Because the survey-specific fields in the WebCASPAR and ABC-SQL Query Tool input SAS data sets had already been mapped to WebCASPAR fields, no further processing of the field data was necessary.

The WebCASPAR data were aggregated by ORC Macro in accordance with the institutional boundaries used in the defining the 200 institutions included in the database.

The SDR is a sample survey of individuals; sample data for individuals were weighted to obtain institution estimates. In addition, because the SDR is a biennial survey that was conducted in odd-numbered years 1987 through 2001, data for even-numbered years 1988 through 2000 were interpolated. Although estimates of the number of Ph.D.s employed at each institution was entered into the database, partially due to concerns about the appropriateness of institutional estimates derived from this data source and partially because it was uncertain in what capacity these Ph.D.s were employed, these data were never used in the analyses.

For some disciplines, NRC scholarly quality ratings data were collected at a more detailed field level than the WebCASPAR fields. Therefore, for these fields, it was necessary to compute a weighted average of the program ratings, and the institution's number of graduates in the program was used to weight the ranking data. Similarly, the NRC scholarly quality rating distinguished indicators were collected at a more detailed field level than the WebCASPAR fields; therefore, for these fields, it was necessary to derive the NRC scholarly quality rating distinguished indicator as follows:

- Institutions rated distinguished in mathematics, are rated distinguished in mathematics/statistics, regardless of the rating in the smaller field,
- Institutions rated distinguished in statistics and that also had a weighted average above 4.00 in mathematics/statistics are counted as distinguished,
- Institutions rated distinguished in three or more of the seven biology subfields are counted as distinguished, and
- Institutions that have a weighted average in biology above 4.00 are counted as distinguished.

B.8 Database Construction, Quality Control, and Documentation

After the data for the 200 U.S. academic institutions to be included in the database were extracted from each data source and processed, the data were stored in a separate, intermediary SAS data set for each data source. These intermediary SAS data sets were then combined into a single SAS data set by merging the records in each intermediary data set by institution FICE, year, and field. After merging the intermediary data sets, rules for converting missing values to zeroes as defined by SRI and NSF were applied to the data. In addition, if an institution did not have data for a particular combination of year and field, a record containing all missing values was generated for that combination of institution, year, and field.

The methods and processes for extracting, transforming, and loading the data included several stages of quality control. First, non-WebCASPAR source data were reviewed for inconsistencies requiring consultation with NSF, SRI, or the data providers. The WebCASPAR source data sets had already been checked extensively against published data and the source data files. The data extracted from WebCASPAR input SAS data sets were spot checked for individual institutions against that institution’s data in WebCASPAR. For the SCI/SSCI data and patents data, totals from the database were compared to totals generated from the source files.

The SAS programs used for extracting, transforming, and loading the data were designed in modules so that if one data source was updated, only the programs and quality control reviews related to that data needed to be repeated. The SAS programs were documented for ease of

maintenance and update. In addition, the intermediary SAS data sets for each data source and the final SAS data set included formats and labels to make the data sets as self-documenting as possible.

The complete system documentation is comprised of four elements:

- Overall systems design with diagrams, charts and text explaining the purpose of database object (i.e., tables, views, SAS format libraries, etc.) and their relationships.
- Technical data model reports (e.g., a code book) with the name, description and columns for each object to include primary and foreign keys, column type and formats, and indexes.
- The programs used to create and update the database and information (e.g., location, structure, access method, periodicity) on the source files.
- The update schedule and methodology to include the sequence in which the programs must be run.

B.9 Conduct Initial Analyses and Direct Further Analyses

After receiving the Publication Trends database, SRI prepared the database for analysis using SAS, JMP, and Excel. This included reformatting and merging of appropriate data, as well as transformations of appropriate analysis variables. For example, as explained later, financial variables were deflated to constant dollars, and various lags were imposed on input variables and citation counts. During this phase of the analysis, SRI also worked with NSF to define field groups.

Typically analyses were conducted first on institutional-level data and then replicated using field group level data. In general, models developed at the institutional level were also found to be appropriate at the field-group level, although the value of the coefficients changed and in some cases there were slight modifications in the model. A factor analysis on the eight different outcome measures revealed that the first principal component captured almost all of the variability. As a result many initial analyses were conducted using the first principal component. Later analyses verified that results using other outcome measures yielded very similar results. As a result, NSF later specified that most analyses be conducted on fractional publication counts for an expanding journal set rather than the first principal component.

Because many of the explanatory variables were highly correlated (i.e., counts of full professors, associate professors, assistant professors, etc.), the model building process involved stepwise regressions on sets of related variables to identify a few variables from each set that appeared to capture the influence of that set on the outcome variables. This was an iterative process, since there was also a high degree of correlation between variable sets (i.e., Academic R&D expenditures and count of Postdoctorates). Due to the large number of observations, variables whose effects were far too small to be practically important were often identified to be statistically significant. Consequently, the criterion for variable inclusion was primarily that a variable increased the model r-squared by an increment of approximately 0.01. A model consisting of 7 variables was developed for the first principal component and it was found that this model required only slight modification for the various publication and citation outcome measures. When the model building process was repeated for field groups, only minor changes

occurred in the model. After the model was developed, SRI and NSF jointly explored the extent to which the model was capable of predicting the change within institutions in publications when resource inputs changed over time. During these analyses residuals were examined to determine if there were explanatory variables that had non-linear effects or subsets of institutions for which the model fit was less satisfactory.

SRI met with NSF to explore the results of the initial modeling at the institutional and field levels. This provided an opportunity to hypothesize explanations for the model regression coefficients and for NSF to direct the analysis by recommending other analyses that they believed would be helpful in interpreting study results or assist them in addressing related issues. SRI conducted follow-up analyses at the institutional and field levels to address recommendations made during the second meeting with NSF or further explore hypotheses generated during that meeting.